基于知识图谱推理驱动的多模态早期中立评估智能体研究

徐麟瑞 [1,2] 李明涛 [3] 王梓轩 [1] 徐炽 [4] 韩林睿 [5*]

1 中国政法大学法治信息管理学院 北京 102249
2 武汉大学信息管理学院 武汉 430072
3 中国政法大学商学院 北京 102249
4 武汉大学计算机学院 武汉 430072
5 中国政法大学数据法治研究院 北京 100088
（linrui_han@163.com）

摘 要 早期中立评估是落实"诉源治理"理念、实现纠纷高效化解的重要机制，但当前面临专业人才匮乏和普及度受限等挑战。大模型技术的发展为破解困境提供了新的可能性，然而其法律逻辑推理能力的不足仍限制了在早期中立评估等法律咨询任务中的应用。本研究一是优化了 LightRAG 知识图谱检索增强技术，提升其可视化法律推理能力和多轮对话记忆功能，使其更适配多轮法律咨询场景，并引入微调后的 ChatGLM3-6B 模型作为问答引擎，同时结合多模态大模型 Lama-3.2-vision 和 SenseVoice-small 分别负责图像与音频处理。二是在《民法典》及其适用解释以及国际仲裁云链线上法律文书数据集的基础上，构建法律知识图谱，并通过数据清洗与提示词工程优化，完成了智能体的集成开发，实现了从多轮法律咨询到证据提交再到早期中立评估文书生成的全流程智能化支持。实验结果表明，与通义千问-Max、通义千问-Turbo、ChatGLM3-6B 和通义法睿-Plus-32k 模型相比，本研究构建的智能体在响应速度和问答质量方面均表现出显著优势。此外，采用模型推理与数据分离范式，确保用户数据仅在本地存储，推理过程加密传输且服务器不存储数据，从而显著提升了个人信息安全性，可为早期中立评估的智能化应用提供有力的技术支撑。

关键词： 早期中立评估；智能体；知识图谱；LightRAG；图模融合

Knowledge Graph Reasoning-Driven Multimodal Early Neutral Evaluation Intelligent Agent

XU Linrui[1,2], LI Mingtao[3], WANG Zixuan[1], XU Chi[4], HAN Linrui[5*]

1 School of Information Management for Law, China University of Political Science and Law, Beijing 102249

2 School of Information Management, Wuhan University, Wuhan 430072

3 Business School, China University of Political Science and Law, Beijing 102249

4 School of Computer Science, Wuhan University, Wuhan 430072

5 The Institute for Data Law, China University of Political Science and Law, Beijing 100088

(email: linrui_han@163.com)

Abstract Early neutral evaluation is a crucial mechanism for implementing the concept of "dispute source governance" and achieving efficient dispute resolution. However, it currently faces challenges such as a shortage of professional talent and limited popularity. The development of large model technology offers new possibilities for overcoming these difficulties, but its insufficient legal logical reasoning ability still restricts its application in legal consultation tasks such as early neutral evaluation. This study, first, optimizes the LightRAG technology, which enhances knowledge graph retrieval, to improve its visual legal reasoning capabilities and multi-turn dialogue memory functions, making it more suitable for multi-turn legal consultation scenarios. It introduces the fine-tuned ChatGLM3-6B model as the Q&A engine and combines the multimodal large model Lama-3.2-vision and SenseVoice-small for image and audio processing, respectively. Second, based on the "Civil Code" and its applicable interpretations, as well as international arbitration cloud chain online legal document datasets, a legal knowledge graph is constructed. Through data cleaning and prompt engineering optimization,

the intelligent agent is integrated and developed, achieving full-process intelligent support from multi-turn legal consultation to evidence submission and early neutral evaluation document generation. Experimental results show that compared with models such as Tongyi Qianwen-Max, Tongyi Qianwen-Turbo, ChatGLM3-6B, and Tongyi Farui-Plus-32k, the intelligent agent constructed in this study demonstrates significant advantages in response speed and Q&A quality. Additionally, by adopting a model reasoning and data separation paradigm, user data is stored only locally, the reasoning process is encrypted and transmitted, and the server does not store data, thereby significantly enhancing personal information security. This can provide strong technical support for the intelligent application of early neutral evaluation.

Keywords Early Neutral Evaluation, Intelligent Agent, Knowledge Graph, LightRAG, Graph-Model Fusion

1 引言

党的二十大以来，“诉源治理”理念被提到新的高度，司法实践对纠纷解决方式提出了更高要求[1]。在此背景下，早期中立评估（Early Neutral Evaluation，简称 ENE）作为替代性争议解决的一种重要方式被引入。早期中立评估是一种旨在诉讼前通过中立第三方的评估来促进纠纷双方达成和解的机制[2]。尽管 ENE 机制在国内已有一些初步尝试，但这些尝试存在可获得性弱、系统化程度低、专业程度良莠不齐等突出问题[2]。传统的人工评估模式已难以适应当前纠纷解决的迫切需求，专家资源的稀缺性与案件数量的激增之间的矛盾日益凸显。

随着大语言模型技术的迅速发展，面向司法领域的生成式人工智能因其在法律语言理解、法律知识问答、法律预测和法律文本生成等领域表现突出，从而在支持法律人进行更高效和精准的司法决策、为公众提供智能化的法律咨询服务等“诉源治理”过程中受到关注。

然而，智能化的 ENE 全流程开发，远不是通过简单引入大语言模型就能够解决的。其一，大多数法律大模型一般是基于通用大模型，通过指令微调和增量训练构建，虽在适应法律任务上有一定的提升，但仍未有效解决生成机制不透明、生成幻觉文本、缺乏解释性等问题；其二，大语言模型无法进行语音输入、证据提交等多模态交互，无法有效满足当事人的多样化案件评估需求；其三，大语言模型在缺乏检索增强(RAG)、结构化思维链提示(CoT)、任务提示和工作流(Pipeline)开发的情况下，难以完成早期中立评估涉及中从法律咨询到法律文书生成的全流程复杂法律任务。

因此，本研究通过引入 LightRAG 知识图谱增强检索技术与智能体(Agent)工作流，提出一种从多轮法律咨询到证据提交，再到早期中立评估文书出具的全流程智能体构建方法。

构建路径如下：第一步，基于法律咨询任务优化后的 LightRAG 知识图谱检索增强技术，结合“夫子·明察”法律微调大模型，构建早期中立评估智能体雏形；第二步，将民商法律知识向量化与结构化处理，经词表清洗后存储到知识图谱当中，以构建智能体的知识图谱检索增强能力；第三步，开发智能体工作流，并引入多模态大模型，支持语音交互和证据提交，优化当事人使用体验；第四步，构建早期中立评估报告模版和基于任务协作的提示词工程，完成全流程早期中立评估智能体的构建。总的看来，本研究的主要贡献有：

（1）基于“夫子·明察”法律大模型，融入调优的 LightRAG 技术，设计并实现了民商案件咨

询法律知识图谱，并用于知识图谱检索增强，优化了大模型的法律逻辑推理能力；

（2）延展了知识图谱与大语言模型的“图模融合”架构，并基于该架构实现了早期中立评估智能体，完成了从多轮法律咨询到证据提交，再到早期中立评估文书出具的全流程智能体的开发，构建了一种智能化 ENE 架构；

（3）通过优化 LightRAG 技术构建了一种模型推理与数据存储分离的模式。数据不出域，所有用户法律数据均保存在本地，只在推理过程中加密传输给服务器用于大模型推理，且服务器不储存任何数据，这保证了用户个人信息的高安全性。

2 相关工作

与本研究的相关工作主要有 AI Agent 领域的进展、大模型推理能力优化研究进展和知识图谱检索增强领域的研究进展。

2.1 AI Agent 领域的研究进展

2024 年 1 月，人工智能领域的华人杰出学者李飞飞的团队对 AI Agent 领域做了系统综述[3]。综述认为，由于大模型具有强大的理解和生成自然语言的能力，它可以作为 Agent 的核心组件，Agent 则可以通过学习大量的文本数据和语言规则，从而进行自主的语言理解和生成。一般的，Agent 被定义为以 LLM 为核心，具备记忆（Memory）、任务规划（Planning）以及工具使用（Action）的集合，其中 LLM 是核心大脑，Memory、Planning 以及 Tool 等则是 Agent 系统实现的三个关键组件，并对每个模块下实现路径进行了细致的梳理和说明。Langchain 公司最新发布的 AI 智能体调查报告中，非技术行业的公司也对 AI 智能体表现出了浓厚的兴趣。90%的非技术公司受访者表示已经或计划将 AI 智能体投入生产，与科技公司的 89%几乎持平[4]。这显示了 AI 智能体的跨行业吸引力：Agent 技术正在进入甚至深刻变革如法律行业这类传统行业。

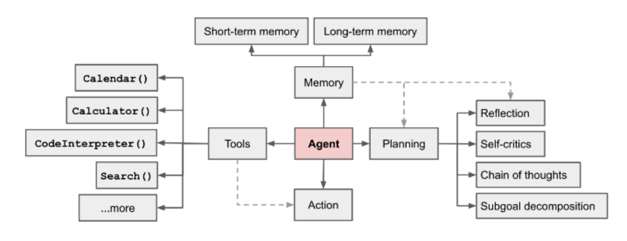


图 1 Agent 的一般架构

Fig.1 General Architecture Diagram of the Agent

2.2 大模型推理能力优化研究进展

大模型推理能力的提升一直以来都是人工智能与自然语言处理领域的一个研究热点。学者们通过各种技术创新和方法探索，已经在多个方面取得了进展。特别是在静态推理、推理服务、结构化思维提示和算法创新方面，学者们已经提出了多种有效的方法来提升大模型的推理能力。此外，跨领域融合和多模态思维链也为大模型推理能力优化提供了新的视角。

总的来看，为提升大模型的逻辑推理能力，已经取得的成果大致可以分为以下三类：（1）推理底层算法的优化；（2）结构化思维模式的建立；（3）辅助技术和算法。

模型推理底层算法的优化关注于提升大模型的推理效率与性能。

其中，静态推理考验的是底层算子的性能，是大模型推理服务优化的基石。通过对计算图改造，并更深入地对算子本身进行优化[5]，静态优化技术能够在不牺牲精度的前提下，提高算力效率，为后续动态推理服务优化奠定基础。静态推理优化方法包括注意力算子优化、线性算子优化以及算子融合的计算图优化。针对注意力和线性算子进行优化能够显著降低大模型推理延时。

动态推理服务是向上承接用户请求，向下调用静态推理的中间层，其优化方法包括批处理技术、调度技术、异步技术及多卡通信技术。对于

算力中心而言，其目标是能用最少的计算存储资源服务最多的用户请求，在满足“服务等级目标”（ServiceLevel-Objective，SLO）的情况下，系统的最大吞吐率（Throughput）的单位是请求/秒（Request/s）。因此，延迟和吞吐的权衡是大模型推理服务面临的主要问题。为了更好地权衡 SLO 和 Throughput，学术界和工业界提出了一系列优化方法，包含以 Orca 为代表的请求批处理方案，以 SplitFuse 为代表的调度技术[6]。同时，调度与执行之间的异步交叠，以及多卡并行也是大模型推理服务的优化技术。

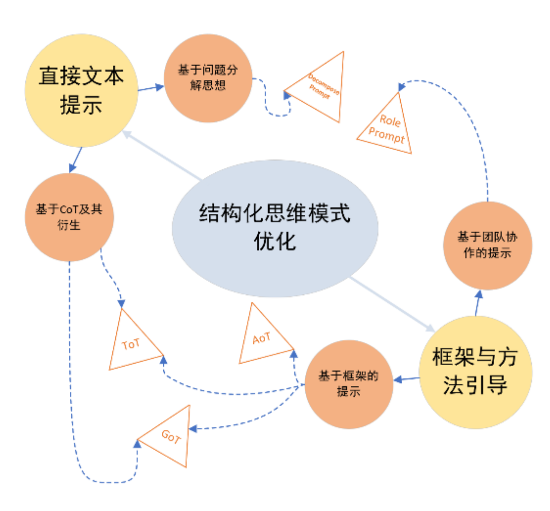


图 2 结构化思维模式优化方法

Fig.2 Structured Thinking Pattern Optimization Methods

结构化思维模式优化方法则从人类认知逻辑的高度，对现有的提示学习范式进行系统性剖析与重构，通过模拟人类解决问题的思维过程来提升大模型的推理能力。如图 2 所示，结构化思维提示包括思维链提示、基于分解思想的提示、基于框架的提示和基于团队协作的提示。[7]从结构化思维提示的角度来看，提示学习大致可以分为两个主要类别。第一类是通过在大语言模型的输入端添加额外文本进行提示。这种方法旨在通过清晰、结构化的提示文本来引导模型的生成过程。这些提示文本通常具有明确的层次结构、逻辑关系和关键词，以帮助模型沿着特定的逻辑路径进行推理与生成。思维链方法就是一种典型的应用，它通过将复杂问题分解为一系列有序步骤，使模型能够逐步推理并得出结论；第二类则是在基础提示文本的基础上，引入额外的框架或方法，以更有效地引导大语言模型完成任务。这种方法借助人类认知理论，采用如 ToT、GoT 等算法来构建解决问题的流程或通过算法框架的导入来指导大模型的推理，如 AoT 的方法[8]。这种框架不仅增强了模型的推理能力，还使其能够处理更复杂的场景。

现有的系列研究给予了本研究很多的启发，尤其是在在模型多轮对话记忆功能设计、早期中立评估文书的出具、证据识别等功能部分。本文将在后文中进行详细论述早期中立评估 Agent 构建中对结构化思维提示的运用和多模型协作框架的运用。

此外，推理辅助技术和方法扮演着至关重要的角色。这些方法通过提供外部辅助和自我评估机制来增强模型的推理能力。首先，分布式技术，尤其是 KV 缓存复用[9]，通过有效利用缓存来节省算力，成为大模型推理优化的关键策略。这种方法允许模型在处理大规模数据时保持高效，从而在不牺牲性能的前提下降低计算成本。这种方法目前已经深度集成到各类大模型的推理架构之中。

其次，角色扮演和工具增强方法通过模拟特定角色的思维方式和提供外部工具及资源，分别增强了模型在特定领域内的推理和决策能力以及处理特定知识或工具需求问题时的灵活性和效率。例如，本文将要讨论的 RAG、GraphRAG[10]以及 LightRAG[11]均为工具增强方法。这些方法使得模型能够更好地适应多样化的任务和环境，提高了其在实际应用中的适应性和实用性。

最后，推理校验技术如 Self-Evaluation[12]和 rStar 方法[13]通过引入自我评估和验证机制来提高模型推理结果的可靠性。这种机制有助于模型在推理过程中发现并纠正错误，确保输出的准确性和可靠性。通过这种方式，模型能够在复杂和不确定的环境中提供更加稳定和可信的推理结果，这对于提高大模型在关键任务中的表现尤为重要。

这些方法都给本研究进行模型推理能力提升以启示：其一，本研究在提升模型推理能力时尝试过 rStar 方法的引入，但由于推理校验所需的时间对于法律咨询任务而言较长，因而舍弃了此方法；其二，本研究通过预设提示词工程、CoT 提示词和 LightRAG 结构化角色数据集来帮助 Agent 理解其早期中立评估咨询师的角色，起到了很好的效果；其三，研究受记忆缓存模式的启发，设计了基于 LightRAG 的长短期记忆模式；最后，本研究的动态推理服务设计了大量的异步函数以提升推理效率。

2.3 知识图谱检索增强技术进展

知识图谱和大型语言模型的融合已经成为了自然语言处理领域的一个热点研究方向。知识图谱能够有效地表示实体间的复杂关系，而 LLM 则在理解和生成自然语言方面展现出了强大的能力。将两者结合，可以使得模型不仅能够理解语言的深层含义，还能够利用结构化的知识进行更准确的信息检索和生成，这对于提升模型在复杂任务中的表现具有重要意义。

GraphRAG 技术正是在这一背景下的技术突破。它由微软研究团队提出，通过构建基于图的文本索引，将实体知识图谱与预生成的社区摘要相结合，使得模型能够针对全局性问题生成全面且多样化的答案。GraphRAG 通过两阶段的索引构建过程，首先从源文档中派生出实体知识图谱，然后为所有密切相关的实体群体预生成社区摘要。在给定问题时，每个社区摘要用于生成部分响应，然后所有部分响应再次汇总以形成对用户的最终响应。这种方法在处理大规模数据集时，相较于传统的 RAG 方法，显示出了在答案全面性和多样性上的显著提升。而在法律咨询任务当中，这也正是当事人所需要的优质回答的需求[14] 。

表 1 LightRAG 与 GraphRAG 技术参数对比

Table 1: Win rates (%) of baselines v.s. LightRAG across four datasets and four evaluation dimensions.

	Agriculture		CS		Legal		Mix	
	NaiveRAG	**LightRAG**	NaiveRAG	**LightRAG**	NaiveRAG	**LightRAG**	NaiveRAG	**LightRAG**
Comprehensiveness	32.69%	67.31%	35.44%	64.56%	19.05%	80.95%	36.36%	63.64%
Diversity	24.09%	75.91%	35.24%	64.76%	10.98%	89.02%	30.76%	69.24%
Empowerment	31.35%	68.65%	35.48%	64.52%	17.59%	82.41%	40.95%	59.05%
Overall	33.30%	66.70%	34.76%	65.24%	17.46%	82.54%	37.59%	62.40%
	RQ-RAG	**LightRAG**	RQ-RAG	**LightRAG**	RQ-RAG	**LightRAG**	RQ-RAG	**LightRAG**
Comprehensiveness	32.05%	67.95%	39.30%	60.70%	18.57%	81.43%	38.89%	61.11%
Diversity	29.44%	70.56%	38.71%	61.29%	15.14%	84.86%	28.50%	71.50%
Empowerment	32.51%	67.49%	37.52%	62.48%	17.80%	82.20%	43.96%	56.04%
Overall	33.29%	66.71%	39.03%	60.97%	17.80%	82.20%	39.61%	60.39%
	HyDE	**LightRAG**	HyDE	**LightRAG**	HyDE	**LightRAG**	HyDE	**LightRAG**
Comprehensiveness	24.39%	75.61%	36.49%	63.51%	27.68%	72.32%	42.17%	57.83%
Diversity	24.96%	75.34%	37.41%	62.59%	18.79%	81.21%	30.88%	69.12%
Empowerment	24.89%	75.11%	34.99%	65.01%	26.99%	73.01%	45.61%	54.39%
Overall	23.17%	76.83%	35.67%	64.33%	27.68%	72.32%	42.72%	57.28%
	GraphRAG	**LightRAG**	GraphRAG	**LightRAG**	GraphRAG	**LightRAG**	GraphRAG	**LightRAG**
Comprehensiveness	45.56%	54.44%	45.98%	54.02%	47.13%	52.87%	51.86%	48.14%
Diversity	19.65%	80.35%	39.64%	60.36%	25.55%	74.45%	35.87%	64.13%
Empowerment	36.69%	63.31%	45.09%	54.91%	42.81%	57.19%	52.94%	47.06%
Overall	43.62%	56.38%	45.98%	54.02%	45.70%	54.30%	51.86%	48.14%

LightRAG 则是在 GraphRAG 的基础上进一步优化的技术，由香港大学的研究团队提出。它在 GraphRAG 基础上通过双级检索系统增强索引类型的区分。在 GraphRAG 中，一切的出发点都是实体（Entity），这是不合理的。 因为抽象的问题，如法律概念的延展性问题，很难找到合适的 Entity。LightRAG 由此提出双层次索引结构——低层次索引：实体为索引的键(Key)，绑定相关的 Entity，Relation，Text，这与 Graphrag 中一致；高层次索引：从实体边中抽象出主题概念，这些主题概念作为为主题键。这种双级检索系统对于法律咨询任务而言是重要的，可以保证模型的回答在尽量避免幻觉的情况下更好地理解抽象法律问题并进行检索。

在 LightRAG 的基础上，可以进一步探索实现模型推理的可视化技术方案。这种方案可以通过构建交互式的可视化界面，让用户能够直观地看到模型是如何利用知识图谱中的实体和关系进行信息检索和生成答案的。通过图形化展示检索

过程和生成步骤，法律咨询者可以更深入地理解模型的内部工作机制，从而提高模型的可解释性和透明度。这种可视化技术方案不仅有助于研究人员分析和优化模型，也为最终用户提供了更直观的交互体验。

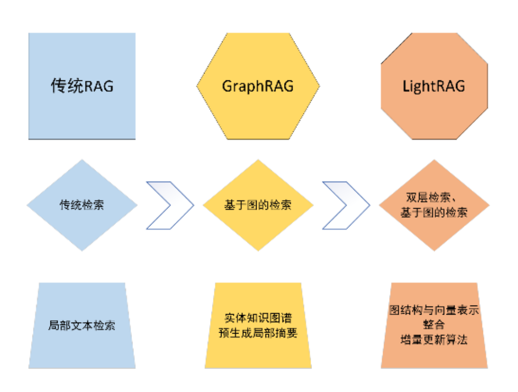


图 3 RAG 技术迭代图

Fig.3 Schema of Evolution in RAG Technology

本研究所构建的早期中立评估 Agent 的核心技术优化就是在 LightRAG 研究的基础上完成的，主要优化包括推理可视化、基于 LightRAG 的多轮长短期记忆、多模态融合和评估文书生成。

3 早期中立评估智能体构建路径

3.1 知识图谱构建

3.1.1 知识实体抽取

知识实体抽取应用了 LightRAG 技术，LightRAG 通过将文档拆分为更小的可管理片段，从而显著提升了检索系统的效率。这种方法使得相关信息的快速识别和获取变得更加便捷，无需逐篇分析整个文档。利用大语言模型来识别和提取各种实体（如人名、日期、地名和事件）及其之间的关系。提取后的信息将用于构建一个综合性的知识图谱，旨在揭示文档集合中的联系与深层见解。[15]

3.1.2 实体关系清洗

数据清洗是数据处理的核心，能确保后续图谱的应用以及数据统计和分析的理想效果。而在实践中，本研究利用 ItInsight 进行数据清洗，旨在消除冗余信息，纠正实体中的错误，并在数据结构中提供一致性，以确保数据的准确性。实体中最常见的错误是拼写错误。拼写错误可以分为谐音拼写错误(如“网洛(络) 功能”) 以及不同的语音拼写错误(如“傅立(里) 叶级数”) 。因此，用大模型来优化整个纠错过程，可以快速纠错，减少人工编辑的工作量和时间。另一个常见错误是语法错误，例如“傅里变换(傅里叶变换) ”等。模糊匹配算法可以编辑实体之间的距离，从而过滤出有语法错误的实体。即对实体字符串进行模糊匹配，计算每个实体与目标实体之间的相似度，将相似度最高的字符串作为目标字符串能够匹配上的实体对象。

3.1.2 知识图谱可视化

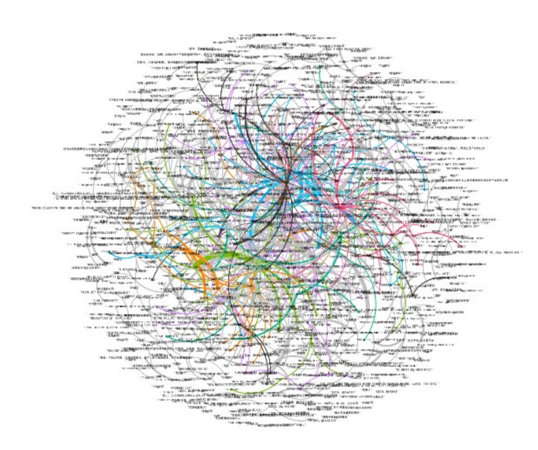


图 4 民商法知识图谱可视化结果

Fig.4 Visualization Results of Civil and Commercial Law Knowledge Graph

本研究采用 ItInsight 以及 Neo4j 图数据库实现民商法知识图谱的高效存储与词表清洗。以 Neo4j 为例，在存储过程中，将知识图谱中的法律实体和关系转换为 Neo4j 中的节点和边，并通过 Cypher 查询语言创建节点和构建关系。随后，利用 Neo4j 的内置可视化工具，直观呈现知识图谱，用户能够清晰浏览法律实体及其关联关系。

如图 4 所示，为《民法典》及其解释的知识

图谱可视化结果，其中的示例数据如表 2 所示：

表 2 示例数据表

源	目标	类型	id	weight	source_id	分组	描述
"中华人民共和国"	"全国人民代表大会"	无向的	0	10	chunk-b7f18b1e19c8ac1525d68bdd0ba88f51	"立法, 法律实施"	"全国人民代表大会在中华人民共和国的领土范围内通过并实施了民法典。"
"民事主体"	"民事权利"	无向的	1	9	chunk-b7f18b1e19c8ac1525d68bdd0ba88f51	"权利保护"	"民事主体享有民事权利，这些权利在民法典中得到保障。"
"民事关系"	"民法"	无向的	2	9	chunk-b7f18b1e19c8ac1525d68bdd0ba88f51	"法律调整"	"民法典调整民事关系，确保民事关系的合法性。"
"民事关系"	"民事责任"	无向的	3	8	chunk-b7f18b1e19c8ac1525d68bdd0ba88f51	"责任追究"	"违反民事关系中的义务可能导致民事责任的产生。"
"民事关系"	"合同"	无向的	4	8	chunk-b7f18b1e19c8ac1525d68bdd0ba88f51	"合同规范"	"合同是民事关系的一种形式，受民法典的规范。"
"民事关系"	"婚姻家庭"	无向的	5	8	chunk-b7f18b1e19c8ac1525d68bdd0ba88f51	"关系调整"	"婚姻家庭关系属于民事关系的一部分，受到民法典的调整。"
"民事关系"	"继承"	无向的	6	8	chunk-b7f18b1e19c8ac1525d68bdd0ba88f51	"关系调整"	"继承关系属于民事关系的一部分，受到民法典的调整。"
"民事关系"	"自然人"	无向的	7	9	chunk-9b952bd955db9432e41bfbfda2144e63	"法律主体"	"自然人是民事关系的主体，包括收养关系和继承关系。"
"民事权利"	"诉讼时效"	无向的	8	7	chunk-b7f18b1e19c8ac1525d68bdd0ba88f51	"时效限制"	"诉讼时效限制了民事权利的行使时间，超过诉讼时效可能丧失胜诉权。"
"民事权利"	"物权"	无向的	9	8	chunk-b7f18b1e19c8ac1525d68bdd0ba88f51	"权利保护"	"物权作为民事权利的一部分，受到民法典的保护。"
"民事权利"	"人格权"	无向的	10	8	chunk-b7f18b1e19c8ac1525d68bdd0ba88f51	"权利保护"	"人格权作为民事权利的一部分，受到民法典的保护。"

3.2 多模态 Agent 构建

3.2.1 语音识模型

本研究使用阿里开源语音模型 SenseVoic-small 帮助实现实时语音识别输入，并使用 Gradio 进行界面设计和模型整合，从而实现了 AI 语音对话功能。上述两个功能的加入大大增加了用户的使用的体验和便捷程度。

3.2.2 图片识模型

在图片识别方面，本研究首先使用了 Llama 3.2-vision 技术实现了图像识别。Llama 3.2-Vision 多模态大型语言模型集合是一个经过预训练和指令调整的图像推理生成模型集合，Llama 3.2-Vision 指令调整模型针对视觉识别、图像推理和回答有关图像的问题进行了优化。

3.3 知识图谱与多模态智能体融合架构

通过将知识图谱中蕴含的法律规则与大语言模型的深度语义理解能力相结合，可显著提升模型的准确性和可解释性，研究的融合架构如图 5 所示。该融合架构通过两种主要方式来增强 Agent 的性能：一是基于 LightRAG 优化的长短期记忆，二是通过 LightRAG 强化模型推理能力。

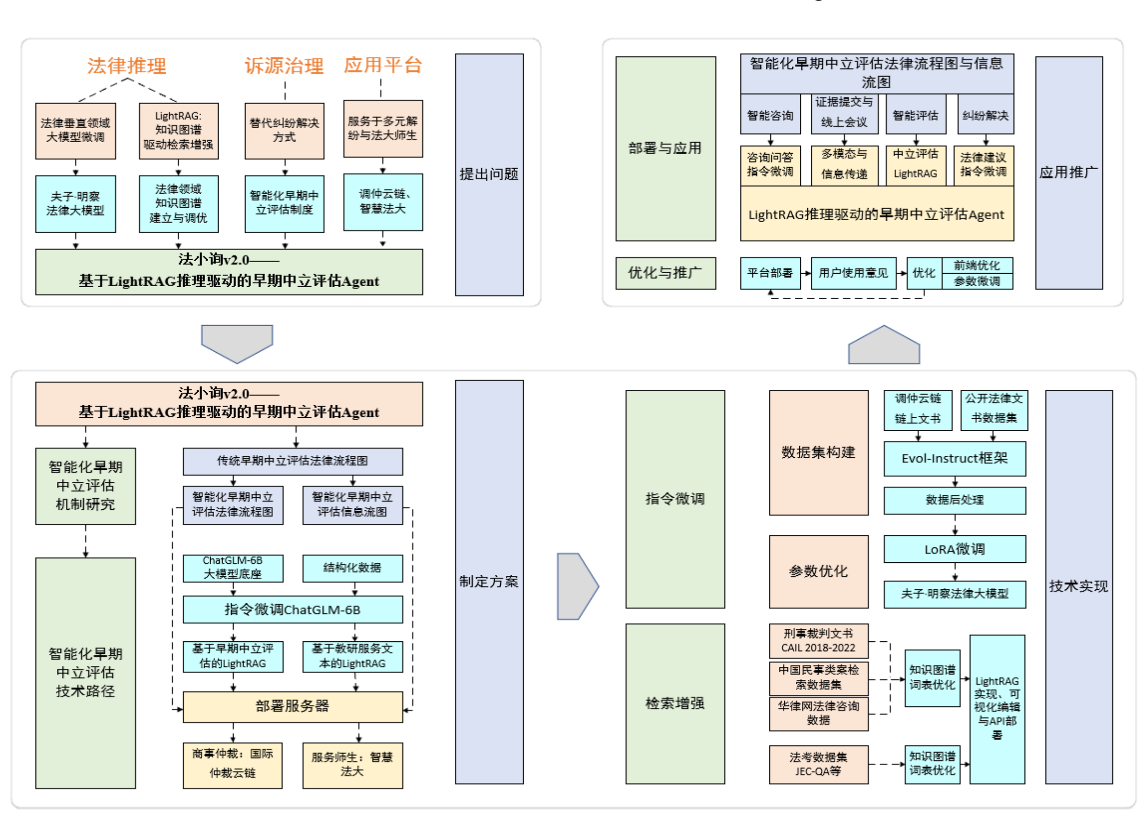


图 5 基于 LightRAG 推理驱动的早期中立评估 Agent 总体技术路线图

Fig.5 A Comprehensive Technical Roadmap for Early Neutral Assessment Agents Driven by LightRAG Inference

3.3.1 基于 LightRAG 优化的长短期记忆

由于在法律咨询任务当中，常会涉及两类场景：第一类是当事人已经较为系统地总结出了案件的来龙去脉，并将其一次性输入到对话框中，这种情况需要模型对当事人的描述进行长期记忆，本研究基于 LightRAG 的增量插入将当事人的案情描述插入到 Agent 的知识图谱当中以实现长期记忆；第二类是当事人第一次对案件进行法律咨询，这种情况下本研究通过主题摘要的方式进行针对对话的短期记忆。

3.3.2 知识图谱检索增强系统

首先，用户输入案件的基本案情信息，系统通过实体识别模块提取关键要素，如涉案主体、涉案客体等。这些实体信息随后在知识图谱中进行链接检索，以获取相关法律条文和案例知识。接着，检索到的知识向量被转化为自然语言，并通过整合至输入模板中，为大模型提供结构化的输入指引。

在中立评估模型中，优化后的提示信息进一步指导模型进行准确的中立评估。大模型不仅生成更为准确的中立评估结果，还能提供每一步分析过程的依据，增强了预测结果的可解释性。最终，系统生成的早期中立评估报告包括相关法律条文与案例、案情分析、诉讼结果预测等，为用户提供可靠的法律建议。

3.3.3 多模态大模型协同工作 Agent

以智能化早期中立评估为例，其包括智能咨询、证据提交与线上会议、线上评估、早期中立评估文书出具四个环节。智能咨询环节，大语言模型对用户提出的自然语言问题做出回答并辅

以语音识别模型实现 AI 语音问答；证据提交与线上会议环节，图像识别模型与 OCR 技术的应用实现图片、PDF、Word、txt 等文件的上传；线上评估环节，大语言模型基于知识图谱将推理可视化，因而模型生成内容可解释性强；最后，早期中立评估文书出具环节，模型自动生成评估文书。通过以上过程，本研究构建了基于案情咨询和证据提交的全流程法律评估。

4. 早期中立评估 Agent 的实现与评估

民事法律咨询知识图谱与大语言模型的融合架构为实现可解释性法律提供了强有力的技术支撑。如图 6，本研究在该架构上进一步融入智能体工作流，该融合架构通过两种主要方式来增强智能体的性能。

首先需要构建的是智能化早期中立评估流程图，并进行数据流动分析。具体流程如图 6 所示。

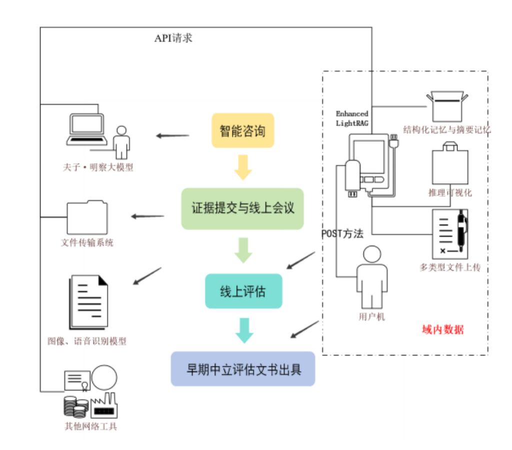


图 6 早期中立评估智能体工作流

Fig.6 Intelligent Early Neutral Evaluation Working Pipeline

民事咨询知识图谱提供了案件相关的法律法规、民商事案件案由认定意见及筛选的案例信息，生成包含法律推理思维链步骤指引的指令模板，明确了民事案件案由确定与法律关系梳理的基本原则、方法及法律裁判可能涉及的法律条款。通过自动化标注与人工校对生成指令微调数据集，并应用 LoRA 策略对大语言模型进行微调。知识图谱提供的丰富结构化信息在微调过程中帮助模型准确理解并应用法律规则。

4.1 全工作流实现

首先，是智能咨询部分。用户输入案件的基本案情信息，系统通过实体识别模块提取关键要素，如民事主体、涉案金额及等。这些实体信息随后在量刑规则知识图谱中进行存储和链接检索，以获取相关法律条文，并实现基于 LightRAG 的

长期记忆。接着，检索到的量刑知识被转化为自然语言，并通过提示组装模块整合至输入模板中，模板包含与案件情节相关的法律条款和量刑规则，为大模型提供结构化的输入指引。

在模型中，通过引入知识图谱的检索增强，大模型不仅生成更为准确的刑期咨询输出，还能提供每一步分析过程的依据，增强了输出结果的可解释性。最终，系统生成基于 LightRAG 的检索回答，并根据提示词工程的提示进行关键性信息补充性发问，来帮助当事人更好地进行下一步的法律咨询。伪代码见算法 1。

证据提交与线上会议部分应用了多模态大模型。用户提交的证据通过 OCR 识别与图像识别大模型识别后存储在本地的结构化知识图谱中，并在本地历史消息中进行备份。该过程服务器不保存用户数据，保证了数据的隐私与安全。线上会议的开展与国际仲裁云链合作进行，基于国际仲裁云链平台进行对接实现。

4.2 效率评估

对于以上两个环节，本研究进行了本地知识图谱推理生成和插入的效率测试实验，实验环境为 12th Gen Intel(R) Core(TM) i5-1240P 1.70 GHz，集成显卡。计算公式参考了 Woosuk Kwon 等人提出的 Normalized Latency 度量[16]：

$$eval_time = \frac{response_time}{input_token * output_token}$$

其中，数据集来自于 CAIL 2024 的法律咨询数据集，共 233 条数据；公式中的 response_time 为 Agent 应答时间，input_token 与 output_token 用于粗略反映问答对的复杂程度，最终得到的 eval_time 用于衡量早期中立评估 Agent 的咨询效率。

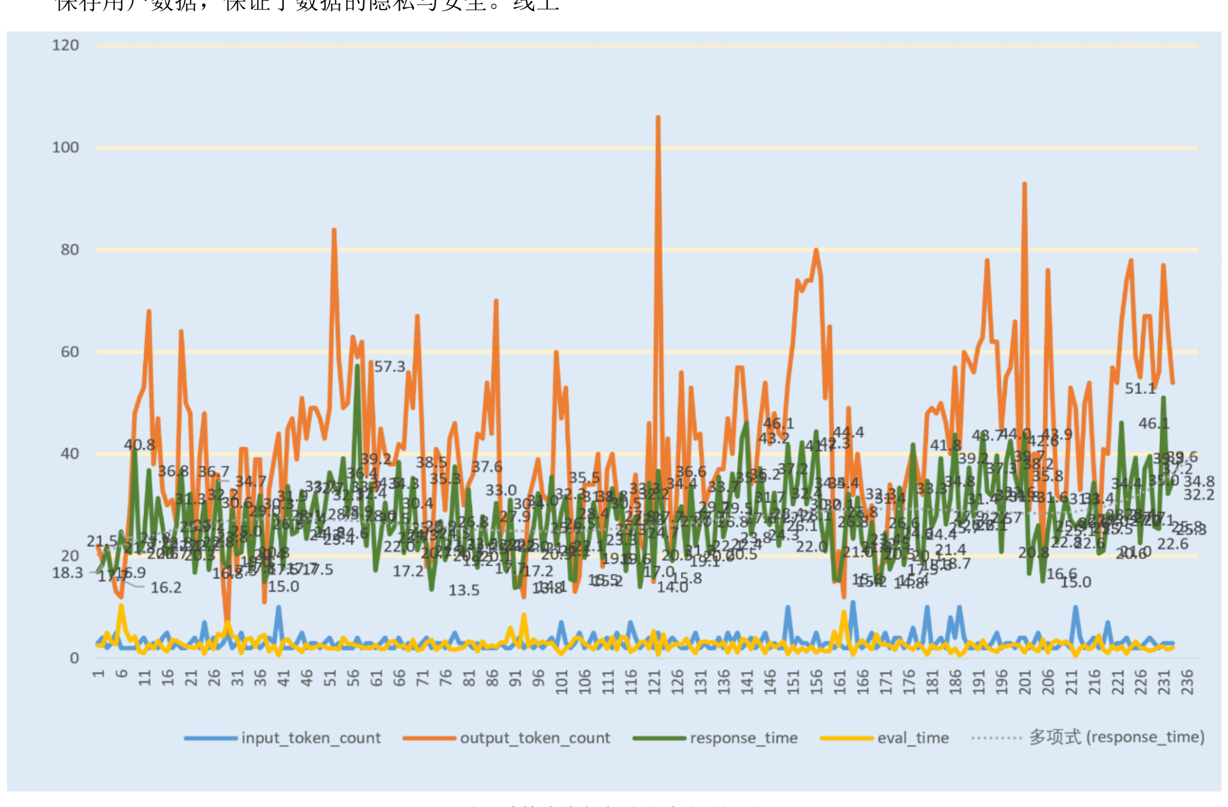


图 7 法律咨询任务综合响应时间图

Fig.7 Legal Consultation Task Response Time Chart

如图 7 所示，大多数的咨询轮次的原始响应时间均在 40s 之下，平均响应时间在 30s 之下；即使是测试集当中最复杂的咨询任务下，输出时间也没有超过 1 分钟。这意味着本研究设计的

Agent 通过 LightRAG 的优化和数据与模型问答分离模式，即使在无独立显卡的普通轻薄级笔记本上运行也可以高效地进行咨询任务。eval_time 反映了 Agent 的效率能力，可以发现其在 input_token 和 output_token 时较少时值较高，说明其核心效率较为稳定，eval_time 这一指标主要受输入和输出 token 数的影响。

算法 1：用户咨询

输入:
- 用户问题 User_Query
- 法律知识库 Legal_Knowledge_Base
- 大语言模型 Fine_Tuned_LLM

输出:
- 法律咨询结果 Legal_Consultation_Result

过程:
1. Entities ← Extract_Entities(User_Query) // 从用户问题中提取法律相关实体
2. Relevant_Legal_Info ← Search_Legal_Info(Legal_Knowledge_Base, Entities) // 在法律知识库中检索相关法律信息
3. NL_Formatted_Info ← Format_Info_Naturally(Relevant_Legal_Info) // 将检索到的法律信息格式化为自然语言
4. Structured_Prompt ← Combine_Info(NL_Formatted_Info, User_Query) // 将格式化信息与用户问题结合，形成结构化提示
5. Legal_Consultation_Result ← Generate_Response(Fine_Tuned_LLM, Structured_Prompt) // 使用大语言模型生成法律咨询结果
6. 返回 Legal_Consultation_Result // 输出法律咨询结果

结束

4.3 生成质量评估

早期中立评估文书的出具通过 LightRAG 案情、检索历史记录反馈、结构化模版和提示词工程实现。可以高效地生成早期中立评估文书。通过将案件信息和相关法律知识整合进预设的模板，系统能够自动提取关键信息并填充到文书的相应部分。提示词工程则确保了文书内容的专业性和准确性，引导系统在生成文书时考虑到民事案由确定原则、物权法和债法二分原则、法律适用范围等关键法律因素。这种自动化的文书生成方式不仅提高了工作效率，还有助于保持文书的一致性和质量，确保每份文书都能反映出案件的核心要素和法律评估。此外，通过不断优化和调整结构化模板和提示词，系统能够适应不同的案件类型和法律环境，提供更加精准和个性化的法律服务。

最终，本研究针对市面上常用的通用大模型通义千问-Max、通义千问-Turbo、ChatGLM3-6B 和通义法睿-Plus-32k 法律大模型与早期中立评估 Agent 的问答对输出内容在阿里云平台做了智能化评估，评估维度包括：

表 3 法律咨询任务生成质量评估维度

评估维度	描述
准确性	确保所提供信息的准确性，以事实为基础，以法律为准绳
相关性	回答必须针对当事人提出的法律问题，避免无关内容
文化敏感与无害	了解和尊重当事人的文化背景和差异，合乎伦理道德
信息丰富性	在保证准确性的同时提供详尽的信息
清晰性	使用清晰、易懂的语言回答问题
当事人参与	鼓励当事人进一步交流，针对案件关键点进行确认与反问
建设性反馈	即使在面对批评性或负面的问题时，也应保持积极和建设性的态度

评估结果如下：

表 4 基于 CAIL 2024 法律咨询数据集的模型问答得分

模型名称	1分	2分	3分	4分	5分	6分	7分	8分	9分	10分	平均分	标准差	中位数
法小询 v2.0	0	0	0	1	16	1	41	174	0	0	7.59	0.84	8.0
通义千问-Turbo	0	0	1	1	33	1	35	162	0	0	7.38	1.11	8.0
ChatGLM3-开源版-6B	0	0	0	6	61	2	24	140	0	0	6.99	1.37	8.0
通义法睿-Plus-32K	0	0	1	2	45	1	33	151	0	0	7.21	1.23	8.0
通义千问-Max-Latest	0	0	0	1	21	2	43	166	0	0	7.51	0.93	8.0

表 4 中可见，除了在前文的实验中 Agent 在相应速度上有出色的表现以外，经“夫子·明察”与经优化的 LightRAG 驱动的早期中立评估 Agent 在法律问答的质量上也超越了一众通用大模型和法律大模型，无论是体现总体问答质量的平均分数据还是体现问答稳定性的标准差数据上，本研究所设计的早期中立评估智能体均（即法小询 v2.0）取得了第一。

4.4 私有化部署与数据安全

本研究通过私有化部署实现了法律咨询服务的本地化处理，确保数据处理的安全性和隐私性。系统在本地服务器上运行，避免了数据在公共网络中的传输，从而减少了数据泄露的风险。特别地，系统通过 VPN 加密通道保护数据传输的安全。咨询与评估界面如图 8 所示。

在数据安全性方面，对于敏感数据，本研究通过采用更强的加密技术 SSH 协议，以确保数据在传输过程中的安全。代码中还包含了对 API 响应状态的检查，确保只有合法的响应才会被处理，这有助于防止恶意数据的注入。通过在代码中设置 API 接口和端口，系统可以控制对敏感数据的访问，只有授权的请求才能访问数据。这些措施共同构成了一个多层次的安全防护体系，保护了敏感数据不被未授权访问和泄露。



图 8 法小询 v2.0 demo 实时运行界面

Fig.8 Fa Xiaoxun v2.0 Demo Real-time Operation Interface

结束语 随着人工智能技术的飞速发展，特别是在法律咨询和服务领域的应用，法律领域见证了从传统人工评估模式向智能化、自动化解决

方案的转变。本研究通过构建基于知识图谱推理驱动的多模态早期中立评估智能体，不仅推动了大模型技术在法律领域的广泛应用，也为法律服务的数字化转型提供了新的思路和工具。

本研究展示了如何通过优化 LightRAG 技术，结合“夫子·明察”法律大模型和多模态大模型 Lama-3.2-vision 与 SenseVoice-small，实现从多轮法律咨询到证据提交，再到早期中立评估文书出具的全流程智能化处理。这一过程有效不仅提高了私有化法律服务的效率和准确性，

在后续的研究中，笔者还希望引入 ART 框架和 Toolformer 方法，使得大语言模型能够进行自动的多步推理并使用工具。ART 会构建一个任务库，并将用户输入与之对比，然后在适当的地方调用工具[17] 。虽然上下文学习可以很好的利用大语言模型的学习能力,但上下文方法相较于微调方法，稳定性相差较多。微调方法 Toolformer[18]则训练语言模型学会自主决定何时调用哪些 API，并将结果融入后续预测，在多个下游任务上实现了零样本性能的大幅提升。

最后，笔者希望本研究能够为同行提供宝贵的经验和启示，共同推动人工智能技术在法律领域的深入发展，为建设更加公正、透明的法治社会贡献力量。同时，笔者也期待与业界专家和学者进行更广泛的交流与合作，共同探索人工智能技术在法律服务中的更多可能性。

参考文献

[1] 姚建军.推进诉源治理满足群众司法新需求[N].人民法院报, 2023-4-17(02 版).

[2] 江和平,黄琪.民商事纠纷中立评估机制的中国发展之路[J].法律适用,2015,(07):22-27.

[3] Durante Z, Huang Q, Wake N. Agent AI: Surveying the Horizons of Multimodal Interaction[J]. arXiv preprint arXiv:2401.03568, 2024.

[4] LangChain. (2024). LangChain State of AI Agents Report. Retrieved from https://www.langchain.com/stateofaiagents

[5] 毛秋力,沈庆飞,李秀红. 面向算力中心的大模型推理优化技术 [J]. 质量与认证, 2024, (09): 40-44.

[6] 陆芸婷,康绍鹏,吴双,等.融合选择核注意力的无纺布缺陷检测[J/OL].计算机工程与应用,1-9[2024-11-26].

[7] 陶江垚,奚雪峰,盛胜利,等. 结构化思维提示增强大语言模型推理能力综述[J/OL]计算机工程与应用, 1-21[2024-11-10].

[8] Suzgun M, Kalai A T. Meta-prompting: Enhancing language models with task-agnostic scaffolding[J]. arXiv preprintarXiv:2401.12954, 2024.

[9] KWON W, LI Z H, ZHUANG S Y, et al. Efficient memory management for large language model serving with PagedAttention[C]//Proceedings of the 29th

[10] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). From Local to Global: A Graph RAG Approach to Query-Focused Summarization. ArXiv, abs/2404.16130.

[11] Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2024). LightRAG: Simple and Fast Retrieval-Augmented Generation. ArXiv, abs/2410.05779.

[12] Ling Z, Fang Y, Li X, et al. Deductive verification of chain-of-thought reasoning[J]. Advances in Neural Information Processing Systems, 2024, 36.

[13] QI Z, MA M, XU J, et al. MUTUAL REASONING MAKES SMALLER LLMS STRONGER PROBLEM-SOLVERS[J]. arXiv:2408.06195v1 [cs.CL], 2024.

[14] 魏斌.法律大语言模型的司法应用及其规范[J].东方法学,2024,(05):57-73.

[15] 黄勃,吴申奥,王文广,等.图模互补：知识图谱与大模型融合综述[J].武汉大学学报(理学版),2024,70(04):397-412.

[16] Kwon W, Li Z, Zhuang S, Sheng Y, Zheng L, Yu C, Gonzalez J, Zhang H, Stoica I. Efficient Memory Management for Large Language Model Serving with PagedAttention[J]. arXiv preprint arXiv:2309.06180, 2023.

[17] Schick T, Dwivedi-Yu J, Dessi R, Railenu R, Lomeli M, Hambro E, Zettlemoyer L, Canceddа N, Scialom T. Toolformer: Language models can teach themselves to use tools[J]. Advances in Neural Information Processing Systems, 2024, 36.

[18] Hao S, Liu T, Wang Z, Hu Z. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings[J]. Advances in Neural Information Processing Systems, 2024, 36.